Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.7, No.4, 2017
www.iiste.org

LOGISTIC MIXED MODELLING OF DETERMINANTS OF INTERNATIONAL MIGRATION FROM THE SOUTHERN ETHIOPIA: SMALL AREA ESTIMATION APPROACH

Tsedeke Lambore Gemecho[1]* (Correspondence author)

[1]* School of Mathematical and Statistical Sciences, Hawassa University, Hawassa, Ethiopia

Ayele Taye Goshu[1]

[1]School of Mathematical and Statistical Sciences, Hawassa University, Hawassa, Ethiopia

Abstract

The main objective of this study is to investigate socio-demographic and economic characteristics of a household on international migration and to estimate small area proportions at district and enumeration area level. Migration status refers to whether a household has at least one member who ever migrated abroad or not. A total of 2288 data are collected from sixteen randomly sampled districts in Hadiya and Kembata-Tembaro zonal areas, Southern Ethiopia. Several versions of the binary logistic mixed models, as special cases of the generalized linear mixed model, are analyzed and compared. The findings of the study reveal that about 39.4% of the households have at least one international migrant, and the rest 60.6% have no such migrants. Based on analysis of the generalized linear model and stepwise variable selection, four predictors are found to be significantly related to household migration status at 5% significance level. These are age, occupation, and educational level of household head and family size. Then twelve mixed models are analyzed and compared. The best fitting model to the data is found to be the logistic mixed regression model consisting of the six predictors with age nested within districts as random effects. Area or district specific random effect has variance of 1.6180. The district level random variation founded on final model with six predictor variables about the presence of migrant in the households such as the variation between districts is 33% and variation within the district is 67%. From analysis of the final model, it is found that the likelihood of a household of having international migrant increases with head's age and family size. An increase of family size by one person increases the log odds of having migrant by 0.131 indicating that large family size is one of the determinants for migration in the study area. The migration prevalence varies among the zones, the districts and the enumeration areas. Household characteristics: age, educational level and occupation of head, and family size are determinants of international migration. Community based intervention is needed so as to monitor and regulate the international migration for the benefits of the society.

Keywords: GLM, GLMM, Migration, Mixed Logistic, Small Area Estimation

1. Introduction

Migration is a complex phenomenon influenced by economic, social, political, geographical and environmental factors. Migration is defined as the movement of a person or a group of

persons, either across an international border, or within a state. It is a population movement, encompassing any kind of movement of people, whatever its length, composition and causes; it includes migration of refugees, internal displaced persons, asylum seekers, smuggled migrants, victims of trafficking, economic migrants, and persons moving for other purposes, including family reunification. It is a large concern for policy makers because flows of population can significantly affects local politics, social, economic, and ecological structures for both sending and receiving countries Chi and Voss, [4] and Abrham, et al., [1].

According to UNDESA, [24] the number of international migrants worldwide has continued to grow rapidly over the past fifteen years reaching 244 million in 2015, up from 222 million in 2010 and 173 million in 2000. Nearly two thirds of all international migrants live in Europe (76 million) or Asia (75 million). Northern America hosted the third largest number of international migrants (54 million), followed by Africa (21 million). Between 2010 and 2015, the international migrant stock grew by an average of 1.9% per year. The majority of the world's migrants live in high-income countries. As of 2015, 71% of all international migrants worldwide equal to 173 million were living in high-income countries. Of these, 124 million migrants where hosted in high-income OECD countries, while 49 million migrants were living in other in high-income non-OECD countries. Only 29% or 71 million of the world's migrants lived in middle or low income countries. Of these, 61 million migrants resided in middle income countries and 9 million in the low income countries. In Africa, Republic of South Africa was the only country hosted the largest numbers of international migrants' equivalent to 3 million in the year 2015.

Rango and Laczko, [19] stated that migration with its associated remittance has diverse socio-economic impacts such as increasing better opportunities for the migrant, improving the livelihood of sending households, contributing economic growth and has emerged as an important policy issue in developing countries. The most recent estimates suggest that there are at least 50 million irregular migrants in the world over one fifth of all international migrants, which is a significant number of whom paid for assistance to illegally cross borders.

The study on the irregular migration of youth from Southern Ethiopia to Republic of South Africa indicates, it is facilitated by a network of human smugglers in Ethiopia work in cooperation with those smugglers from Kenya and Somalia (Teshome, et al., [22]). The problem of irregular migration to Republic of South Africa is widely observed in two zones of the Southern Ethiopia, namely in Hadiya and Kembata-Tembaro Zones. The study results on quantitative cross-sectional study, which was carried out on the randomly selected 4 local districts of two zones. The study revealed that irregular migration was denominated by young aged 20-34 and the conclusion made indicates that most of the young adults who move illegally to Republic of South Africa had suffered several problems like being smuggled, physical abuse, and human right violation and in some cases even death (Teshome, et al., [22]). It is known that at regional and national level of Ethiopia many people of Hadiya and Kembata Tembaro zones are migrating to the Republic of South Africa. The households residing in the two zones are sending young adults irregularly to Republic of South Africa and elsewhere abroad are explored.

The main objective of this study is to investigate impacts of socio-demographic and economic characteristics of a household on international migration and to estimate small area proportions at district and enumeration area level. The specific objectives are to: evaluate the socio--demographic and economic characteristics of migrant and non-migrant households in districts of Hadiya and Kembata Tembaro zones; estimate the local district and enumeration area level proportions of international migration; and develop generalized linear mixed models for international migration status. The study is conducted in highly vulnerable areas by irregular migration. Results are expected to be used as a basis for planning, decision and policy makers and different program implementation at the regional as well as national level in Ethiopia. This study can be a basis to conduct in-depth further studies in specific aspects of international migration along with small area estimation techniques.

2. Methodology

1.1. Description of the Study Area

The study areas are Hadiya and Kembata Tembaro zones which are highly vulnerable areas by irregular migration in Southern Ethiopia. Based on statistical report of the 2007 population and housing census results Hadiya Zone has a total count of 231,846 households and Kembata-Tembaro Zone has a total count of 122,580 households. Hadiya and Kembata Tembaro zones have 11 and 8 districts respectively.

1.2. Sampling Design

In this study the multi-stage sampling design is employed as the sampling design. When the number of small areas is large, it is not feasible for travel cost or time to survey some units in all of them. For travel cost or time to survey, it is sometimes more convenient to use a multi-stage sampling design. Therefore, the sample can be made from administratively clustered small areas and often reduces interviewer travel costs. Sample design for small areas can be determined by only surveying a subset of small areas. In such case sample designs represented by multi-stage sampling where clusters are considered as small areas (Molefe, [16]; Longford, [13]). In this study the sampling frames of 19 local districts as small areas. Four-stages sampling technique are implemented; in the first stage sample of local districts is taken as the primary sampling units. After selecting a sample of local districts, the second stage is selection of samples of Kebeles within each selected local districts and third stage is sampling of enumeration areas and fourthly households within the each selected enumeration areas are chosen. A re-listing of all households in sampled enumeration areas are carried out as suggested by Levy and Lemeshow, [11].

1.3. Sample Size Determination

A total population of 354,426 households is grouped exclusively and exhaustively into 19 local districts in such a way that each small area contains a number of Kebeles, enumeration areas and households as subpopulations. Multi-stage sampling is used: first 16 local districts are selected, at second stage 71 Kebeles are selected, followed by the selection of 89

enumeration areas in the third stage, finally using systematic random sample selection 2381 households are selected. The sample size determination for local districts is used to estimate the proportions of international migration using the formula (Cochran, [5]; Levy and Lemeshow, [11]; Naing, et al., [17]):

$$n_c = \frac{N_c}{1 + N_c d^2} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

Where, N_c is the total number of local districts or clusters and d level of precision $d = 0.1$. This gives the calculated number of local districts, $n_c = 16$. Using the same procedures from a total of 328 Kebeles and 511enumeration areas, 71 Kebeles and 89 enumeration areas are sampled.

The total number of households in the selected 16 local districts is N, which is equal to 301,531. Then the sample size required n_0 of households is the first estimated by equation (2) with finite population correction in equation (3):

$$n_0 = \frac{\sum_{h=1}^{n_c} W_h P_h(1 - P_h)}{d^2 \Big/ Z_{\alpha/2}{}^2} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3)$$

where P_h the proportion of international migration in each small area is taken as 0.5. W_h is the proportion of the population in each local district $h = 1,2, \dots, n_c = 16$ computed as ratio of subpopulation size N_h to the total size N, $Z_{\alpha/2} = 1.96$ is a critical value of the standard normal distribution with significance level $\alpha = 5\%$, and level of precision $d = 0.02$. Then using equation (2), the sample size is estimated to be $n_0 = 2401$ and after finite population correction by (3), it becomes $n = 2381$ households.

1.4. **Description of the Data**

Data on international migration are collected from households using designed questionnaire. At each household level the interview is carried out with the household heads by well trained numerators. Data collections at local district level are coordinated by Woreda labour and social affairs officers. At each enumeration areas level data are collected by selecting one enumerator with good performance from Kebele agricultural extension workers, health care workers and Kebele administration heads.

The response variable is the migration status of a household as reported by household heads (HH). It is a dichotomous with outcome value, $y_{ij} = 1$, if there is at least one migrant in the household ever migrated abroad and if not the outcome value, $y_{ij} = 0$. Predictor variables focus on socio-demographic and economic characteristics of each household and its head: gender of head, age of head, marital status of head, educational status of head, place of residence, family size, occupation of head, ethnicity of head, religion of head, farm land size, zone, district, and enumeration area of household. There is about 4% non-response rate and so final data size accessed from 2288 households. Therefore, a total of 2288 household head data

are collected from 16 randomly sampled local districts and 86 enumeration areas are collected respectively.

1.5. Statistical Model

1.5.1. Small Area Estimation Methods

Sample surveys have long been recognized as cost-effective means of obtaining data on wide-ranging topics of interest at frequent intervals over time. In most surveys, estimates are used in practice to provide estimates not only for the total population of interest but also for a variety of subpopulations (Rao, [20]; Lohr, [12]). Small area estimation is the process of using statistical models to link survey outcome variables to a set of predictor variables known for small areas, in order to predict area-level estimates. It is becoming important in survey sampling due to a growing demand for reliable small area statistics from both public and private sectors. Small area estimation method seeks to improve the precision of the estimates when standard methods are not accurate enough and produces estimations for the small areas having not reliable direct estimators (Rao, [20]; Pfeffermann, [18]; Setiawan and Tarumi, [21]). In this study, small area estimation technique is used to predict area level estimates for proportions of international migration.

1.5.2. Generalized Linear Mixed Model

Extension of linear models to generalized linear models (GLM) was first proposed by McCullagh and Nelder, [15] by noting that the linear model consists of three components: (i) independent observations (ii) mean of observation as linear function of some covariates, and (iii) constant variance of observation. The observation has probability distribution that belongs to the exponential family. The variance is a function of the mean of observation. GLMs generalize a variety of models including normal, binomial, Poisson, and multinomial. In GLMs, the predictor variables **x** affect the response **Y** via the linear predictor. The GLM is obtained by specifying some function of the response conditional on the linear predictor and on other parameters.

Another important extension of GLM is the generalized linear mixed model (GLMM). In the GLMM, the linear predictor contains both fixed and random effects and can be applicable to several areas (Jiang, [10]; Faraway, [6]; McCulloch and Searle, [14]; Zhao, [25]). The response is a random variable; **Y** follows an exponential family distribution defined as:

$$f(y|\theta,\phi) = exp\left\{\frac{y\,\theta - \mathrm{b}(\theta)}{a(\phi)} + c(y,\phi)\right\} \dots\dots\dots\dots\dots\dots\dots (4)$$

where $\mathrm{b}(\theta)$, $\mathrm{a}(\phi)$ & $\mathrm{c}(y,\phi)$ are known functions and ϕ is a dispersion parameter which may or may not be known. The expectation of the response variable $E(Y) = \mu$ and the linear predictor are linked using a link function $g(\mu)$ given fixed effects parameters $\boldsymbol{\beta}$ and random effects $\boldsymbol{v}$ which can be expressed generally as:

$$g(\mu) = \mathbf{X}'\beta + \mathbf{Z}'\upsilon \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (5)$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices of predictors. A very special case of GLMM is mixed logistic regression model with the response variable having Bernoulli distribution in the exponential family and the logit link function $g(\mu) = \text{logit}(\mu)$.

Area level models relate small area direct estimators to area-specific covariates. Rao, [20] and Pfeffermann, [18] considered sampling models, which was originally studied for small area estimation by involving direct survey estimators and linking model for the small area parameters of interest.

1.5.3. Mixed Logistic Regression Model

Suppose Y_{ij} is the binary response variable of interest, where $Y_{ij} = 1$ if there exists at least one international migrant in a household and $Y_{ij} = 0$ otherwise for each cluster $i = 1,2,\dots,n_c$ and household $j = 1,2,\dots,n_i$. Here n_c is number of clusters and n_i is number of households within cluster i. The success probability is defined by $\mathrm{P_{ij}} = \mathrm{Prob}(\mathrm{Y_{ij}} = 1|\text{random effects})$ and $[\mathrm{Y_{ij}} \mid \mathrm{P_{ij}}] \overset{\text{iid}}{\sim} \text{Bernoulli}(\mathrm{P_{ij}})$. The parameters of interest are the small area proportions $\hat{P}_i = \sum_j y_{ij}/n_i$ for each cluster $i = 1,2,\dots,n_c$. We consider the mixed effects logistic regression model (Pfeffermann, [18]; Jiang, [10]; Rao, [20]; Fay-Herriot, [7]), which is a special case of the generalized linear mixed model (GLMM). It is defined as follows:

$$\text{logit}(\mathrm{P_{ij}}|\mathbf{U},\boldsymbol{\upsilon}) = \beta_0 + (\beta_1 + \mathrm{U_1})\mathrm{X_{ij1}} + (\beta_2 + \mathrm{U_2})\mathrm{X_{ij2}} + \cdots + (\beta_s + \mathrm{U_s})\mathrm{X_{ijs}} + \beta_{s+1}\mathrm{X_{ij(s+1)}} + \cdots \beta_k\mathrm{X_{ijk}} + \upsilon_{0i} + \varepsilon_i \dots\dots (6)$$

where β_0 is fixed intercept term; X_{ijr}, $r = 1,2,\dots,k$ are household level k covariates; β_r, $r = 1,2,\dots,k$ are fixed regression coefficients; U_r, $r = 1,2,\dots,s < k$ is a random effect due to $\mathrm{X_{ijr}}$ with $\mathrm{U_r} \overset{\text{iid}}{\sim} N(0,\sigma_u^2)$; and $\upsilon_{0i} \overset{\text{iid}}{\sim} N(0,\sigma_{\upsilon_0}^2)$ are area specific random effects for the cluster $i = 1,2,\dots,n_c$. The total data size is $n = \sum_{i=1}^{n_c} n_i = 2288$. The clusters manly represent districts, but may also represent enumeration areas. Several versions of the model (6) are analyzed.

1.5.4. Parameter Estimation Methods

The logistic regression model may be estimated by using either full maximum likelihood (ML) or a GLM methodology. Maximum likelihood estimation typically uses modified forms of Newton–Raphson estimating equations; GLM uses an iteratively re-weighted least squares (IRLS) algorithm that is a simplification of maximum likelihood estimation but is limited to distributions belonging to the exponential family of distributions. In the case of maximum likelihood, an estimating equation is defined as setting to 0 the derivative of the log-likelihood function of the response distribution with respect to one of the parameters of interest, where there is a single estimating equation for each unknown parameter (Hilbe, [8]). The IRLS algorithm and related statistical values are based on the formula for the exponential family of distributions in equation (4). The term $c(y,\phi)$ is the normalization term, which is required to

assure that the probabilities sum to 1. For the logistic model, as well as the Poisson and negative binomial count models, the scale is taken as 1. The first derivative of b (θ) with respect to θ, or b′(θ) is the mean; the second derivative (b″(θ)) is the variance. These are extremely useful relationships and are not found in other distributions. Changing the link gives the user alternate models. For instance, the logit link $\ln(\mu/(1-\mu))$ is the natural link for the binomial distribution (Faraway, [6]; McCulloch and Searle, [14]; Hilbe, [8]).

GLM applications typically come with a variety of goodness of fit statistics, residuals, and so forth, to make the modeling process much easier than traditional ML. In fact, this is the advantage of using GLM methods over individual ML implementations. Alternatively, for the logistic model, the ML algorithm can provide easier use of so-called Hosmer–Lemeshow fit statistics, which are based on collapsing observations having the same pattern of covariates. The likelihood associated with the mixed models for binary data considered in equation (5) is:

$$L(\gamma, \sigma^2_{u0}|\mathbf{y}, \mathbf{X}, \mathbf{Z}) = \prod_j \int_{-\infty}^{+\infty} \prod_i g(y_{ij}|X_{ij}, Z_j, u_{0j})\, f(u_{0j}) du_{0j}, \ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

where

$$g(y_{ij}|X_{ij}, Z_j, u_{0j}) = \mu_{ij}^{y_{ij}}(1-\mu_{ij})^{1-y_{ij}},$$

$$\mu_{ij} = 1 - F\left(-\left\{\gamma_{00} + \sum_{p=1}^{P}\gamma_{p0}x_{pij} + \sum_{q=1}^{Q}\gamma_{q0}z_{qj} + u_{0j}\right\}\right),$$

$$f(u_{oj}) = \frac{1}{\sqrt{2\pi}\sigma_{u0}}\exp\left(\frac{u_{oj}^2}{2\sigma_{u0}^2}\right)$$

Statistical computing programming language R evaluates the integral $L(\gamma, \sigma^2_{u0}|\mathbf{y}, \mathbf{X}, \mathbf{Z})$ for the binary response model using standard Gaussian quadrature or adaptive Gaussian quadrature for the numerical integration (Hilbe, [8]; Berridge and Crouchley, [2]). There is not an analytic solution for this integral with normally distributed u_{oj}. The analyses are made in R software version 3.3.2 and SPSS version 20.

1.5.5. **Intra-class Correlation Coefficient**

For binary data, the intra-class correlation coefficient is often expressed in terms of the correlation between the latent responses Y^*. The logistic distribution for the level-one residual, ε_{ij}, implies a variance of $\pi^2/3 = 3.29$ (Berridge and Crouchley, [2]; Browne, et al., [3]). This means that, for a two-level random intercept model with an intercept variance of σ^2_{u0}, the intra-class correlation coefficient is:

$$\rho = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \pi^2/3} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

3. Results and Discussions

3.1. Descriptive Statistics

The main objective of this study is to investigate impacts of socio-demographic and economic characteristics of a household on international migration and to estimate small area proportions at district and enumeration area levels of Hadiya and Kembata-Tembaro zones, Southern Ethiopia. Primary data are collected with a sample survey conducted from July 2016 — October 2016 for the purpose of PhD study on international migration status in Southern Ethiopia. Out of 2381 sampled households, data are obtained from 2288 households' of 16 districts capturing 86 enumeration areas. Out of 2288 households 65.6% and 34.4% are interviewed from Hadiya and Kembata Tembaro zones respectively.

The response variable is migration status (1 if there is at least one member of the household ever migrated abroad, 0 otherwise). Predictor variables focus on demographic and socio-economic characteristics of each household and household head: sex of head, age of head, marital status of head, place of residence, family size, educational status of head, occupation of head, ethnicity of head, religion of head, farm land size, zone, district and enumeration area.

The proportions of migrant households those had at least one person ever migrated abroad is 39.4% and the rest 60.6% of households have no international migrants at their home. Out of 902 migrant households the proportion of 68.6% are found in Hadiya zone and 31.4% are found in Kembata-Tembaro zones. The proportions of 41.2% and 36% of interviewed households are migrant households within Hadiya and Kembata Tembaro Zones correspondingly.

More than 30% proportions of migrant households observed in 11 districts are listed in descending order such as: Misha (72.1%), Angacha (72%), Lemo (61%), Damboya (57.1%), Hossana town (55.9%) & Anna Lemo (52.3%), Doyogena (40.5%), Gibe (34.4%), Duna (34.3%), Soro (30.5%) and Kacha Bira (33.3%). Less than 30% proportions of migrant households observed in 5 districts are listed in descending order such as: Shashogo (26.1%), Hadero Tunto (24.4%), Durame Town (14.6%), Kedida Gamela (8.3%), and Misraq Badawacho (5.4%).

Table 1. International Migrant Proportions at District Level

Zone Name	District Name	Migrant Count (%)	No Migrant Count (%)	Total Count (%)
Hadiya	1. Misha	137 (72.1)	53 (27.9)	190 (12.7)
	2. Gibe	55 (34.4)	105 (65.6)	160 (10.7)
	3. Lemo	100 (61.0)	64 (39.0)	164 (10.9)
	4. Shashogo	42 (26.1)	119 (73.9)	161 (10.7
	5. Misraq Badawacho	7 (5.4)	122 (94.6)	129 (8.6)
	6. Soro	84 (30.5)	191 (69.5)	275 (18.3)
	7. Duna	60 (34.3)	115 (65.7)	175 (11.7)
	8. Anna Lemo	58 (52.3)	53 (47.7)	111 (7.4)
	9. Hossana Town	76 (55.9)	60 (44.1)	136 (9.1)
Sub total		619 (41.2)	882 (58.8)	1501 (100.0)
Kembata Tembaro	10.Angacha	88 (71.0)	36 (29.0)	124 (15.8)
	11.Doyogena	45 (40.5)	66 (59.5)	111 (14.1)
	12.Damboya	52 (57.1)	39 (42.9)	91 (11.6)
	13.Kacha Bira	50 (30.3)	115 (69.7)	165 (21.0)
	14.Kedida Gamela	10 (8.3)	111 (91.7)	121 (15.4)
	15.Hadero Tunto	33 (24.4)	102 (75.6)	135 (17.2)
	16.Durame Town	6 (14.6)	35 (85.4)	41 (5.2)
Sub total		283 (36.0)	504 (64.0)	787 (100.0)
Grand Total	Grand Total	903 (39.4)	1386 (60.6)	2272 (100.0)

Out of 86 enumeration areas 32 had more than 50% proportions of migrant households' within each enumeration area. Also 32 enumeration areas with more than 50% proportions of migrant households are found within the 11 districts mentioned above by observing more than 30% of migrant households.

The proportion of presence of migrant in the households relating to age composition of household heads are 64.3%, 47.3%, 37.6%, 28.8%, and 27.6% for the age categories ≥ 60, 50-59, 40-49, 31-39 and 19-30 respectively. The result shows that household heads within older age had more proportion of international migrants. The marital status of most of the household heads (89.44%) is married and the rest 10.66% of household head marital status are single & divorced/widowed. Concerning the educational status of household heads, the proportion shows that the prevalence of migration decreases with increasing educational level of the heads. About 77.9% of the respondents reside in the rural areas and the rest in urban areas. Equal proportions (39.4%) of migrant households are observed in both rural and urban resident households. Large proportion of respondent household heads (77.9%) religion is Protestant followed by 12.1% Orthodox, 6.6% Muslim, 2.1% Catholic and 1.3% others.

The proportion of interviewed household heads with their respective ethnic groups are dominated by two ethnic groups-Hadiya and Kembata. These two ethnic groups account the highest share of the respondent household heads of 61.5% belongs to ethnic group -Hadiya followed by 31.1% of ethnic group-Kembata. The rest ethnic groups of respondent household heads are Donga, Amhara, Guraghe, Silete, Dubamo and others together shares 7.4%.

The largest proportions of occupational distribution of household heads are engaging in farming tasks with the proportion of 69.7% followed by 10.1% of merchant. The rest 20.2% of respondents are student, housewife and others. The average and standard deviation of family size are 6.32 and 2.29 respectively. Family sizes in the range 5-8 had more proportion of migrant households while compared to less than 4 and more than 9 family sizes.

Table 2. Descriptive Statistics for Selected Predictor Variables

Predictor Variables	Categories	Migration Status				Total
		Migrant HH	Percent	Non-migrant HH	Percent	
Sex of HH	Male	745	38.70%	1180	61.30%	1925
	Female	156	43.20%	205	56.80%	361
	Total	901	39.40%	1385	60.60%	2286
Age of HH	19-30	64	27.60%	168	72.40%	232
	31-39	145	28.80%	359	71.20%	504
	40-49	332	37.60%	550	62.40%	882
	50-59	195	47.30%	217	52.70%	412
	≥ 60	164	64.30%	91	35.70%	255
	Total	900	39.40%	1385	60.60%	2285
Educational status of HH	Can't read/write	318	38.70%	504	61.30%	822
	Can read/write	349	42.50%	472	57.50%	821
	Uneducated	667	40.60%	976	59.40%	1643
	Primary School(1-8)	133	37.70%	220	62.30%	353
	High School(9-12)	66	37.70%	109	62.30%	175
	Higher Education	33	29.20%	80	70.80%	113
	Educated	232	36.20%	409	63.80%	641
	Total	899	39.40%	1385	60.60%	2284
Marital Status of HH	Single	25	37.30%	42	62.70%	67
	Married	805	39.30%	1243	60.70%	2048
	Divorced/Widowed	70	41.20%	100	58.80%	170
	Total	900	39.40%	1385	60.60%	2285
Occupation of HH	Government Employee	61	33.90%	119	66.10%	180
	Farmer	609	38.20%	984	61.80%	1593
	Merchant	98	42.60%	132	57.40%	230
	Student	24	42.90%	32	57.10%	56
	House Wife	69	48.90%	72	51.10%	141
	Other	39	45.90%	46	54.10%	85
	Total	900	39.40%	1385	60.60%	2285
Ethnic group of HH	Hadiya	556	39.50%	852	60.50%	1408
	Kembata	273	38.30%	439	61.70%	712
	Guraghe	20	66.70%	10	33.30%	30
	Silete	4	33.30%	8	66.70%	12
	Dubamo/Denta	3	30%	7	70%	10
	Donga	13	23.60%	42	76.40%	55

	Amhara	23	65.70%	12	34.30%	35
	Other	10	38.50%	16	61.50%	26
	Total	902	39.40%	1386	60.60%	2288

3.2. **Bivariate Analysis Results**

The association tests show that there is significant association between the response household migration status and the predictors: age, ethnicity, occupation of head, family size, district, and enumeration area at 5% significance level. Its association with sex of head, educational level of head, and farm land size are significant at 10% significance level. However, insignificant associations are found between the response migration status with religion, marital status, and place of residence of household head.

3.3. **Assessing Model Fit**

The overall significance is tested, which is derived from the likelihood of observing the actual data under the assumption that the model has been fitted is accurate. The deviance is the log-likelihood of the final model to the log-likelihood of a null model with no predictor variables. The deviance between –2*log-likelihood for the final model is 2865.18 and for the null model is 3056.22. Therefore, the full model gets smaller deviance, which is good fit to the dataset.

The presence of relationship between the response and combination of predictor variables are based on the statistical significance of the final model chi-square. In this analysis, the distribution reveals that the probability of the model chi-square (χ^2 (17)) with value 191 was $2.2x10^{-16}$, which is less than 5% level of significance. The null hypothesis that there was no difference b/n null and final model was rejected. Therefore, the final model predicts the response variable well and it is good fit to the data.

The analogous to the linear regression coefficient of determination, R^2 have been proposed for the logistic regression are Cox & Snell R^2, Nagelkerke $R^2$and the McFadden R^2 value, they provide an indication of the amount of variation in the response variable. In linear regression, R^2 has a clear meaning; it is the proportion of the variation in the response variable that can be explained by predictor variables in the model. Attempts have been developed to yield an equivalent of this concept for the logistic model. However, it renders the meaning of variance explained for the logistic regression. The pseudo R^2 value of 6.25% (McFadden pseudo-R^2) and 8.04% (Cox & Snell and Nagelkerke R^2) indicates that the inclusion of the predictor variables in the model reduces the variation as measured by absolute value of log-likelihood of null model.

The overall accuracy of the final model to predict migration status of household is 65% correctly predicted. And 85.6% of absence of migrant and 31.3% of presence of migrant in household are correctly predicted in their respective categories. The Hosmer–Lemeshow test that yields a χ^2 (8) value of 8.851 and is insignificant with p-value of 0.335 which is greater than 5% significance level. This suggesting the final model is good fit to the data well. In other words, the null hypothesis, Ho: null model is a good fit to data is reasonable rejected.

The adequacy of the fitted model is checked for possible presence and treatment of outliers and influential observations. The minimum and maximum values of the test results for Cook's influence statistics are 0.0054 and 0.05127, respectively. DFBETAs for model parameters and Cook's influence statistics are both less than unity, which shows that an observation had no overall impact on the estimated vector of regression coefficients. There is no observations have Studentized residuals less than -3 or greater than +3, then we can conclude that there are probably no outliers in the dataset. Therefore, we can go on to evaluate and interpret the model parameters.

3.4. **Fixed Effects Logistic Regression Analysis**

The classical logistic regression model constitutes fixed effects only and is defined as:

Model0 $$\text{logit}(P_{ij}) = \beta_0 + \beta_1 \text{Age}_{ij} + \beta_2 \text{Job}_{ij} + \beta_3 \text{Educ}_{ij} + \beta_4 \text{Hsize}_{ij} + \varepsilon_{ij} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (9)$$

We use the algorithm of variable selection suggested by Hosmer and Lemeshow, [9]. The algorithm involves variable selections decision at each step of the modeling process. First, fit a univariate model with each of the covariates. Second, select more candidates that are significant at some chosen significance level to build a multivariate model. Any variable whose univariable test has a p-value less than 0.25 is a candidate for the multivariable model along with all variables of known intuitively relevant variables regardless of their statistical significance. Third, following the fit of the multivariable model, the importance of each variable included in the model should be verified. Fourth, once we have obtained a model that we feel contains the essential variables, we should look more closely at the variables in the model. Fifth, once we have refined the main effects model and ascertained that each of the continuous variables is scaled correctly, we check for interactions among the variables in the model.

Using the algorithm and forward-backward variable selections the some categories of four predictor variables such as age, occupation, and educational level of household head (HH), and family size are found statistically significant at 5% significance level. The predictor variable ethnic group of HHs is entered into final model as dummy variables without variable selection as ethnic1 (1=Hadiya, 0=others) and Ethnic2 (1=Kembata, 0=others). In the Table 3, the analyzed results of regression coefficients, standard error, z-value, p-value, odds ratio and 95% CI of odds ratio using logistic regression analysis for fixed effects are displayed.

Table 3. Results of Fixed Effects Logistic Regression Analysis

Parameter	Estimate	Std. Error	z value	Pr(>|z|)	Odds Ratio	95% CI OR
β_0(intercept)	-1.902	0.325	-5.848	4.98e-09***	0.149	(0.078, 0.281)
β_{12}(AgeHHs, 19-30) Ref						
β_{12}(AgeHH, 31-39)	0.029	0.182	0.156	0.876	1.029	(0.722, 1.476)
β_{13}(AgeHH, 40-49)	0.427	0.171	2.501	0.01238*	1.532	(1.101, 2.152)
β_{14}(AgeHH, 50-59)	0.808	0.188	4.302	1.69e-05***	2.244	(1.559, 3.258)
β_{15}(AgeHH, ≥ 60)	1.663	0.213	7.822	5.18e-15***	5.277	(3.494, 8.047)

β_{21}(JobHHs, Gov. Employee) Ref						
β_{22}(JobHH, Farmer)	0.016	0.203	0.080	0.93608	1.016	(0684, 1.519)
β_{23}(JobHH, Merchant)	0.383	0.236	1.623	0.10461	1.467	(0.925, 2.334)
β_{24}(JobHH, Student)	0.458	0.344	1.334	0.18231	1.581	(0.802, 3.338)
β_{25}(JobHH, House Wife)	0.716	0.265	2.703	0.0069*	2.047	(0.804, 3.094)
β_{26}(JobHH, Other)	0.622	0.295	2.108	0.03504*	1.863	(1.219, 3.450)
β_{31}(Educ can't read/write) Ref						
β_{42}(EducHH read/write)	0.281	0.110	2.558	0.01053*	1.324	(1.068, 1.643)
β_{43}(EducHH Primary 1-8)	0.200	0.142	1.417	0.15642	1.222	(0.925, 1.613)
β_{44}(EducHH High 9-12)	0.212	0.188	1.130	0.25836	1.236	(0.853, 1.783)
β_{45}(EducHH Higher Educ)	-0.246	0.267	-0.921	0.35690	0.782	(0.460, 1.312)
β_4(Family size)	0.123	0.0205	6.012	1.83e-09***	1.131	(1.087, 1.178)

We can determine which predictor variables matter in logistic regression analysis by looking at the P-values of the individual coefficients. Predictor variables with P-values that are less than 5% significance level would be considered as statistically significant. Meaning that there is statistical evidence that they affect the probability that the response variable is 1, which is the presence of international migrant in the household. More generally, if the P-value is less than α, then a predictor variable is statistically significant at α level of significance. Therefore, the categories of the predictor variables identified by star(s) are statistically significant at 95% confidence level. But there is no statistical evidence to the categories of predictor variables matter on the migration status of household their respective P-value is greater than 0.05 and no interpretation is made for these.

For instance, the p-value of age of household head older than 60 years is $5.18e^{-15}$. Thus there is strong statistical evidence that household heads are more likely to send household members abroad if they are older more than 60 years. In general the results in the age categories of HHs illustrate as age of heads increases they are more likely to send household members abroad. The p-value of family size (Hsize) is $1.83e^{-09}$. Thus there is strong statistical evidence that household heads are more likely to send household members abroad if family size increases by a person.

The z-value is the regression coefficient divided by its standard error. If the z-value is large in magnitude (that is, either positive or negative), it indicates that the corresponding true regression coefficient is not 0 and the corresponding predictor-variables matters on the response variable. A good rule of thumb is to use a cut-off value of 2 which approximately corresponds to a two-sided hypothesis test with a significance level of α=0.05. For instance, for the occupation of household head in categories of farmer, merchant, student are having the z-values 0.079, 1.627 and 1.339 which are not large enough to provide strong evidence to be significant.

Odds ratio (OR) - is a measure of association between an exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the

odds of the outcome occurring in the absence of that exposure. In this study, the response variable migration status denotes the presence or absence of international migrant and one of the predictor variable AgeHHs denotes the age of household head with categories 19-30, 31-39, 40-49, 50-59, and older than 60 years. For instance, the odds ratio of age of HHs older than 60 is 5.277 estimates that presence of migrant is 5.277 times more likely to occur among household head with this age group than among the reference age category 19-30. Also, for the quantitative predictor variable family size, the odds ratio shows the presence of international migrant in the household increase by 1.131 for every a person increase in household. We can follow the same procedures to interpret the rest results in the Table 3.

3.5. **Small Area Estimation of Binomial Proportions**

Small- area estimation refers to estimation of parameters for a large number of geographical areas when each has relatively few observations. For instance, one might want district or enumeration area-specific estimates of characteristic such as the proportion of households having one or more international migrants. Small area estimation models are random effects models. These models treat each small area as a cluster with its own random effect coming from a common distribution of the random effects.

Table 4. Estimated Proportions of Households having Migrants

Districts	N_i	n_i	n_p	π_i	Fixed Effects		Random effects		
					Pred prob	*pred logit*	*Pred prob*	*pred logit*	u_i
Misha	24,033	190	137	0.15	0.433	-0.288	0.712	0.972	1.532
Gibe	20,071	160	55	0.06	0.387	-0.488	0.344	-0.684	0.0267
Lemo	20,804	164	100	0.11	0.427	-0.309	0.596	0.421	0.996
Shashogo	20,804	161	42	0.05	0.362	-0.597	0.263	-1.087	-0.281
Misraq Badawacho	27,166	129	7	0.01	0.350	-0.651	0.068	-2.719	-1.884
Soro	34,529	275	84	0.09	0.367	-0.579	0.306	-0.875	-0.099
Duna	22,000	175	60	0.07	0.389	-0.476	0.343	-0.686	0.0274
Anna Lemo	13,887	111	58	0.06	0.463	-0.159	0.517	0.074	0.581
Hossana Town	16,962	136	76	0.08	0.432	-0.298	0.553	0.239	0.801
Angacha	15,581	124	88	0.1	0.425	-0.326	0.699	0.915	1.302
Doyogena	13,920	111	45	0.05	0.407	-0.409	0.405	-0.414	0.008
Damboya	14,404	91	52	0.06	0.365	-0.409	0.557	0.262	0.867
Kacha Bira	20,499	165	50	0.06	0.373	-0.560	0.305	-0.884	-0.31
Kedida Gamela	15,316	121	10	0.01	0.377	-0.523	0.101	-2.296	-1.786
Hadero Tunto	17,063	135	33	0.04	0.376	-0.532	0.248	-1.179	-0.513
Durame Town	4,960	40	6	0.01	0.384	-0.513	0.173	-1.695	-1.14

Let denote the finite population size by N and assume that it is partitioned into 16 non-overlapping districts (or small areas), each of sizes N_i with i=1, 2, 3… 16 for districts such that $N = \sum_{i=1}^{16} N_i$. In this study, one of the limitations is the true proportions of migrant households in each district level are not found. This is due to the absence of exact number of

migrant households in each district. We have only the number of migrant households that considered in conducted sample survey.

In Table 4, π_i is the probability that $Y_{ij} = 1$ and the values under this column is the proportion of migrant household compared to total sample size at each district level. The result in column n_i is the number sample size drawn from each district and n_p is the number of migrant households at each district. Let u_i be the random area effect for the district i and the random effect results, u_i is each district level variation in migration status of households.

Predicted response probability and predicted logit for each district are explored (see Table 4). The values under predprob indicates the predicted probabilities that can be revalidated with the actual outcome to determine the predicted probabilities indeed associated with the presence of migrant in the household at each districts. It predicts the logit of presence of migrant in household from a set of predictors at each district level. Moreover, it is the predicted probability of presence of migrants in the household at district level.

3.6. **Logistic Mixed Regression Analysis**

3.6.1. **Random Intercept Models**

In a two-level model we split the residual into two components, corresponding to the two levels in the data structure. We denote the district-level residuals called district random effects, by u_j and the household residuals by e_{ij}. The two level extension of which allows for district effects is given by $y_{ij} = \beta_0 + u_j + e_{ij}$, where β_0 is the overall mean of y (across all districts) and $u_j \sim N(0, \sigma_\upsilon^2)$. Three null models with random intercepts defined below as versions of model (6). These are models for district and enumeration area level effects on migration status of household. The inputs are as defined in model (6).

$$\textbf{Model0Zone} \qquad \text{logit}(P_{ij}) = \beta_{0zone} + \upsilon_{i,zone} + \varepsilon_{ij} \qquad (8)$$

$$\textbf{Model0Dist} \qquad \text{logit}(P_{ij}) = \beta_{0dis} + \upsilon_{i,dist} + \varepsilon_{ij} \qquad (9)$$

$$\textbf{Model0EA} \qquad \text{logit}(P_{ij}) = \beta_{0ea} + \upsilon_{i,ea} + \varepsilon_{ij} \qquad (10)$$

3.6.2. **Null Model with District Level Random Effects**

The assumption of random effects with zero mean and constant variance is attained, that is, $\upsilon_{i,dist} \sim (0, 1.022)$. The intercept is interpreted as the log-odds that y = 1 when x = 0 and $\upsilon = 0$ and is referred to as the overall intercept in the linear relationship between the log-odds and x. The log-odds of presence of international migrant in household an 'average' at district level (with $\upsilon_{0,j} = 0$) is estimated as $\hat{\beta}_0 = -0.5827$. This indicates that the overall estimated mean of migration status (across districts) is -0.5827. The mean for district j is estimated as $-0.5827 + \upsilon_{0j}$, where the variance of υ_{0j} is estimated as $\sigma^2_{\upsilon 0} = 1.022$. The district level random variation, $\sigma^2_{\upsilon 0} = 1.022$ and the logistic distribution for the level-one residual, εij, has a variance of $\pi^2/3 = 3.29$. Therefore, the intra-class correlation coefficients,

$\rho = \frac{\sigma^2_{\upsilon 0}}{\sigma^2_{\upsilon 0}+\sigma^2_{\varepsilon}} = 0.237$ indicates that there is 23.7% of the variation between the districts and the rest 76.3% variation is within the district. Also the correlation between randomly chosen pairs of individuals belonging to the same district is 0.237.

3.6.2.1. Testing for District Effects

To test the significance of district effects, we can carry out a likelihood ratio test comparing the null model with a null single-level model. The null model takes the form $\text{logit}(P_{ij}) = \beta_0$. The likelihood ratio statistic for testing the null hypothesis, that is, $H_0: \sigma^2_{\upsilon 0} = 0$, can be calculated by comparing the null model, with the corresponding null single-level model without the random effect. The likelihood ratio test statistic is calculated as two times the difference in the log likelihood values for the two models. The likelihood ratio test statistic is 320.06 with 1 DF, so there is strong evidence that the between-district variance is non-zero.

3.6.2.2. Examining Districts Effects

To estimate the district-level residuals $\hat{\upsilon}_{0j}$ and their associated standard errors, we use the function `ranef` in R with the `condVar` option. This creates a random effects object containing the variance-co-variance matrix in the `condVar` attribute. The 16 district level residuals are stored in υ_0 and $\upsilon_0[1]$ is the list corresponding to the first set of random effects. The estimates of the district effects, $\hat{u}_{0j}$ by obtaining from the null model are examined. To calculate the residuals and produce a 'caterpillar plot' with the district effects in rank order together with 95% confidence intervals we can use the function `ranef()` that can be work for the continuous and categorical responses of two-level random intercepts model. There is only one set of random effects; the `postVar()` attribute only contains the "posterior variance" of each district-level residual. To access this set of variances, we look into the attribute `postVar()` of the data frame $\upsilon_0[[1]]$. This returns a three-dimensional array with the third dimension referring to each individual residual. The district residuals and their standard errors have been calculated and stored for each individual district. We can therefore calculate summary statistics and produce graphs based on these data.

Table 5. District Level Residual and Its Standard Error

District	Residuals	Std. Error
Mis/Badawacho	-2.014	0.328
Kedida Gamela	-1.661	0.296
Durame Town	-1.003	0.385
Hadero Tunto	-0.525	0.196
Shashogo	-0.444	0.176
Kacha Bira	-0.243	0.167
Soro	-0.235	0.130
Duna	-0.066	0.157
Gibe	-0.061	0.164
Doyogana	0.194	0.190
Anna Lemo	0.651	0.187
Hossana Town	0.797	0.170
Damboya	0.816	0.208
Lemo	1.005	0.158

Angacha	1.424	0.192
Misha	1.496	0.159
$\upsilon_{i,dist} \sim (0, 1.022)$		

The district *Misraq Badawacho* had an estimated random effect residual of -2.014. For district *Misraq Badawacho* the estimate mean migration status is -0.5835 – (-2.014) = 1.441. In contrast, the mean for district 16—Misha is estimated as -0.5835 + 1.496= 0.912. Finally, we use the plot and segments commands to produce a 'caterpillar plot' to show the district effects in rank order together with 95% confidence intervals.

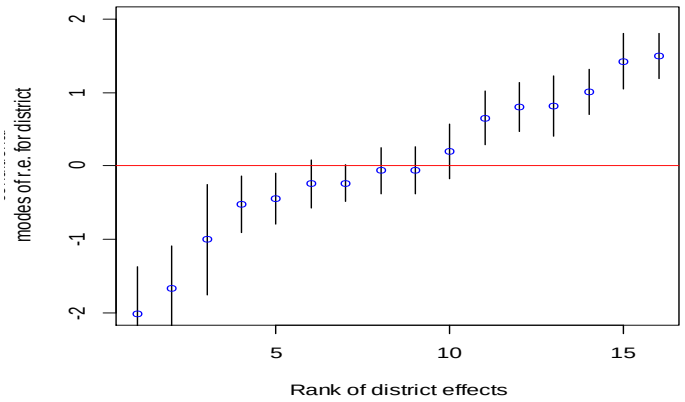

Figure 1. Plot of Estimated Random Effects for Districts

The plot in Figure 1 is the estimated residuals for all districts in the sample. The 95% confidence interval does not overlap the horizontal line at zero, indicating that presence of international migrant in the districts are significantly above the average or below the average.

3.6.3. Null Model with Enumeration Area Effects

Fitting the null model that allows random effects for enumeration area on migration status of household computed using R. The log-odds of presence of migrant in an 'average' enumeration area (one with $\upsilon_{0ea} = 0$) are estimated as $\hat{\beta}_{0,ea} = -0.5751$. This indicates that the overall estimated mean of migration status (across EAs) is -0.5751. The mean for enumeration area ea is estimated as $-0.5774 + \upsilon_{0ea}$, where the variance of υ_{0ea} is estimated as $\hat{\sigma}^2{}_{\upsilon ea} = 1.659$. The enumeration area level random variation, $\sigma^2{}_{\upsilon ea} = 1.659$ and the logistic distribution for the level-one residual, ε_{ij}, has a variance of $\pi^2/3 = 3.29$. Therefore, the intra-class correlation coefficients, ρ is equal to 0.335 indicates that there is 33.5% of the variation between the enumeration areas and the rest 66.5% variation is within the enumeration area. Also the correlation between randomly chosen pairs of individuals belonging to the same district is 0.335.

3.6.3.1. Testing for Enumeration Area Effects

To test the significance of enumeration area effects, we can carry out a likelihood ratio test comparing the null model with the null single-level model. The likelihood ratio statistic for testing the null hypothesis, that is, H_0: $\sigma^2{}_{\upsilon ea} = 0$, can be calculated by comparing the null

model, with the corresponding null single-level model without the random effect enumeration area. The likelihood ratio test statistic is calculated as two times the difference in the log likelihood values for two models. The likelihood ratio test statistic is 361.58 with 1 DF, so there is strong evidence that the between-enumeration area variance is non-zero.

To estimate the enumeration-level random effects or residuals, $\hat{v}_{0i}$ and their associated standard errors, we use the `ranef` () R function with the `condVar` () option. This creates a random effects object, containing the variance-covariance matrix in the `condVar` () attribute. The 86 enumeration area level random effects residuals are stored in v_{0i}. The estimates of the enumeration area effects or residuals, $\hat{u}_{0i}$ obtained from the null model are examined through the following procedures. The residuals and producing a 'caterpillar plot' with the enumeration area effects with 95% confidence intervals can be produced using two-level random intercepts model in R.

Table 6. Estimated Enumeration Area Residual and Its Standard Error

Enumeration Area	Residuals	Std. Error
Mb024_03	-2.427282118	0.7192199
Hd004_01	-2.376847206	0.7251072
Kd009_01	-2.294525890	0.7350391
Kd01_02	-2.234388725	0.7425595
Kd01_05	-2.234388725	0.7425595
.	.	.
.	.	.
.	.	.
M004_01	1.936818202	0.4279178
Ang014_03	2.129702668	0.4512904
M028_02	2.239323271	0.4877586
HS002_09	2.942713872	0.6814753
Lm019_03	3.212081895	0.6547809

We use the plot and segments commands in R to produce a 'caterpillar plot' to show the enumeration area effects in rank order together with 95% confidence intervals.

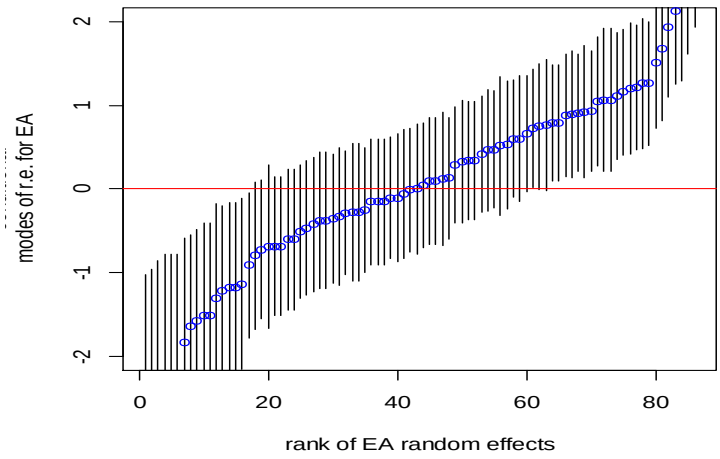

Figure 2. Plot of Estimated Random Effects for Enumeration Areas

The plot in figure 2 is the estimated residuals for all enumeration areas in the sample. For a substantial number of enumeration areas, the 95% confidence interval does not overlap the

horizontal line at zero, indicating that the presence of international migrant in the enumeration areas are significantly above the average or below the average.

3.7. **Mixed Logistic Regression Model with Covariates**

3.7.1. Comparison of Models

Three models (8)—(10) are analyzed in the section 3.6. Here, the following nine GLMM models constructed from model (6) are analyzed. A model with best fit to the data is determined and further analyzed. These models involve covariates and random effects.

Model1 $\text{logit}(P_{ij}) = \beta_0 + \beta_1 Age_{ij} + \beta_2 Job_{ij} + \beta_3 Educ_{ij} + \beta_4 Hsize_{ij} + \beta_5 Ethnic1_{ij} + \beta_6 Ethnic2_{ij} + \upsilon_{i,dist} + \varepsilon_{ij}$ ……(11)

Model2 $\text{logit}(P_{ij}) = \beta_0 + \beta_1 Age_{ij} + \beta_2 Job_{ij} + \beta_3 Educ_{ij} + \beta_4 Hsize_{ij} + \beta_5 Ethnic1_{ij} + \beta_6 Ethnic2_{ij} + U_1 + \varepsilon_{ij}$ ……(12)

Model3 $\text{logit}(P_{ij}) = \beta_0 + \beta_1 Age_{ij} + \beta_2 Job_{ij} + \beta_3 Educ_{ij} + \beta_4 Hsize_{ij} + \beta_5 Ethnic1_{ij} + \beta_6 Ethnic2_{ij} + U_1 + U_2 + \varepsilon_{ij}$ ……(13)

Model4 $\text{logit}(P_{ij}) = \beta_0 + \beta_1 Age_{ij} + \beta_2 Job_{ij} + \beta_3 Educ_{ij} + \beta_4 Hsize_{ij} + \beta_5 Ethnic1_{ij} + \beta_6 Ethnic2_{ij} + U_1 + U_2 + \upsilon_{i,dist} + \varepsilon_{ij}$ ……(14)

Model5 $\text{logit}(P_{ij}) = \beta_0 + \beta_1 Age_{ij} + \beta_2 Job_{ij} + \beta_3 Educ_{ij} + \beta_4 Hsize_{ij} + \beta_5 Ethnic1_{ij} + \beta_6 Ethnic2_{ij} + U_1 + \upsilon_{age:i} + \varepsilon_{ij}$ ……(15)

Model6 $\text{logit}(P_{ij}) = \beta_0 + \beta_1 Age_{ij} + \beta_2 Job_{ij} + \beta_3 Educ_{ij} + \beta_4 Hsize_{ij} + \beta_5 Ethnic1_{ij} + \beta_6 Ethnic2_{ij} + \upsilon_{age:i} + \varepsilon_{ij}$ ……(16)

Model7 $\text{logit}(P_{ij}) = \beta_0 + \beta_1 Age_{ij} + \beta_2 Job_{ij} + \beta_3 Educ_{ij} + \beta_4 Hsize_{ij} + \beta_5 Ethnic1_{ij} + \beta_6 Ethnic2_{ij} + \upsilon_{i,zone} + \varepsilon_{ij}$ ……(17)

Model8 $\text{logit}(P_{ij}) = \beta_0 + \beta_1 Age_{ij} + \beta_2 Job_{ij} + \beta_3 Educ_{ij} + \beta_4 Hsize_{ij} + \beta_5 Ethnic1_{ij} + \beta_6 Ethnic2_{ij} + U_1 + U_2 + \upsilon_{i,zone} + \varepsilon_{ij}$ ……(18)

Model9 $\text{logit}(P_{ij}) = \beta_0 + \beta_1 Age_{ij} + \beta_2 Job_{ij} + \beta_3 Educ_{ij} + \beta_4 Hsize_{ij} + \beta_5 Ethnic1_{ij} + \beta_6 Ethnic2_{ij} + U_1 + \upsilon_{age:i,zone} + \varepsilon_{ij}$ ……(19)

Using the both forward-backward stepwise variable selection techniques the predictor variables namely, age, occupation, and educational level of household head and family size are found to be the significant at 5% level of significance. The predictor variables ethnic groups are included as dummy variables in the final model without variable selection techniques. Then final model that includes: age, occupation, and educational level of household heads, family size, ethnic1 and Ethnic2. Then 9 models are proposed from model (6) including 6 predictors and random effects. The model comparison results in Table 7 illustrate those models: 1, 4, 5 and 6 are found to be nearly equal AIC values. The -2*log-likelihood values of these models have almost similar results. However, the AIC value model 6 is smaller than the rest of models. Therefore, model 6 is selected as the final model and the interpretation of regression coefficients of predictors in this model are done.

Table 7. Results for Model Comparison

Model	AIC		-2*log-likehood	
	AIC	DF	Values	DF
Model0ea	2700.2	2	2696.228	2
Model0dist	2739.6	2	2735.615	2
Model0zone	3059	2	3055.04	2
Model 1	2622.9	14	2594.859	14
Model 2	2903	14	2874.988	14
Model 3	2905	15	2874.988	15
Model 4	2626.9	16	2594.859	16
Model 5	*2624.8*	*29*	*2566.824*	*29*
Model 6	*2622.8*	*28*	*2566.824*	*28*
Model 7	2893.8	14	2865.847	14
Model 8	2897.8	16	2865.847	16
Model 9	2917.6	29	2859.62	29

The results for mixed model analysis of Model 6 are given in Table 8. The age of head, occupation of head, educational level of head, and family size are significant at 5% level of significance. As age of head increase the odds of a household having migrant increase with reference to the lowest age group 19-30. With reference to occupation as government employee, the odds of having migrant is different for households with heads in merchant, housewife and other job categories. Educational level of head is also found important predictor of migration status. The odds of having migrant in a household with a head who can read/write are different from the head who can't read/write. However, it is not significantly different for heads with educational level of primary, secondary and higher education, indicating that these households might behave similarly in terms of sending their members to the international migration. Family size is significant and affecting the odds of migration positively. The odds of a household have migrant increases with family size. Ethnicity is not significant in this case. There exist random effects due to district level and age of head within districts.

Table 8. Analysis Results of the Final GLMM Model

Parameter	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
Fixed effects:				
Intercept	-2.12293	0.51160	-4.150	3.33e-05*
Age 19-30 Ref				
Age 31-39	-0.05702	0.26751	-0.213	0.831216
Age 40-49	0.15780	0.27101	0.582	0.560384
Age 50-59	0.60003	0.32697	1.835	0.06649*+
Age ≥60	1.29761	0.34054	3.810	0.000139*
Job Gov Employee Ref				
Job Farmer	0.21114	0.22773	0.927	0.353858
Job Merchant	0.83369	0.26343	3.165	0.001552*
Job Student	0.35715	0.37401	0.955	0.339623
Job Housewife	1.01943	0.29221	3.489	0.000485*

Job Other	0.69827	0.32245	2.166	0.030349*
Educ can't read/write Ref				
Educ Read/Write	0.21000	0.12539	1.675	0.09399*+
Educ Primary sch (1-8)	0.04160	0.15862	0.262	0.793098
Educ High sch (9-12)	0.11946	0.20903	0.572	0.567657
Educ Higher education	-0.29492	0.29965	-0.984	0.325009
Family size	0.13005	0.02285	5.690	1.27e-08*
Ethnic1 Hadiya	-0.15318	0.21508	-0.712	0.476332
Ethnic2 Kembata	0.12703	0.31184	0.407	0.683757
Random effects:	Variance	Std. Dev.		
District level $\sigma^2_{\upsilon_0}$	1.6180	1.2720		
Age 19-30 Ref				
Age 31-39 $\sigma^2_{u_{11}}$	0.2797	0.5289		
Age 40-49 $\sigma^2_{u_{12}}$	0.3118	0.5584		
Age 50-59 $\sigma^2_{u_{13}}$	0.6864	0.8285		
Age ≥60 $\sigma^2_{u_{14}}$	0.6008	0.7751		

*significant at 5% significance level, *+significant at 10% significance level

The fitted model for each household within district (omitting indices for convenience) is:

$$\log\left\{\frac{\hat{P}}{1-\hat{P}}\right\} = -2.12293 - 0.05702\ \text{Age}(31-39) * I_{age} + 0.1578\ \text{Age}(40-49) * I_{age} + 0.60003\ \text{Age}(50-59) * I_{age} + 1.29761\ \text{Age}(>60) * I_{age} + 0.21114\ \text{JobFarmer} * I_{job} + 0.83369\ \text{JobMerchant} * I_{job} + 0.35715\ \text{JobStudent} * I_{job} + 1.01943\ \text{JobHousewife} * I_{job} + 0.69827\ \text{JobOther} * I_{job} + 0.21\ \text{Educ Readwrite} * I_{edu} + 0.0416\ \text{Educ Primary School}(1-8) * I_{edu} + 0.11946\ \text{Educ High School}(9-12) * I_{edu} - 0.29492\ \text{Educ Higher Education} * I_{edu} + 0.13005\ \text{family size} - 0.15318\ \text{Ethnic Hadiya} + 0.12703\ \text{Ethnic Kembata} + \hat{\upsilon}_0$$

This fitted model can be used to make point estimates given information of individual household. The predication of probabilities or proportions of individual households in each district of having international migration can be made using $\hat{P} = \exp(fitted)/(1+\exp(fitted))$. The predicted probabilities of each district are made as shown in the Table 4 and the interpretation of the regression coefficients can be made as done in Table 7 for fixed effects. The district level random variation with covariates is, $\sigma^2{}_{\upsilon 0} = 1.618$ and the logistic distribution for the level-one residual, ε_{ij}, has a variance of $\sigma^2_{\varepsilon_{ij}} = \pi^2/3 = 3.29$. Therefore, the intra-class correlation coefficients, ρ is equal to 0.33 indicates that there is 33% of the variation between the districts and the rest 67% variation is within the district.

4. Conclusions

The main objective of this study is to investigate impacts of socio-demographic and economic characteristics of a household head and household on international migration and to estimate small area proportions at district and enumeration area level.

A total of 2288 data are collected from sixteen randomly sampled districts in Hadiya and Kembata-Tembaro zonal areas, Southern Ethiopia. The response variable migration status refers to whether a household has at least one member who ever migrated abroad or not. The

findings of the study reveal that about 39.4% of the households have at least one international migrant, and the rest 60.6% have no such migrants. Proportions of households within districts to have international migrants are estimated. Based on analysis of the generalized linear model and forward-backward stepwise variable selection four predictors are found to be significantly related to household migration status at 5% level of significance. These are age, occupation and educational level of household head and family size.

Several versions of the generalized linear mixed models are proposed, analyzed and compared. The best fitting model to the data is found to be the logistic mixed regression consisting of the six predictors with age nested within districts as random effects. Under this model, the district specific random effect is significant with variance of 1.6180. From analysis of the final model, it is found that the odds of a household head of having international migrant increases with head's age and family size. An increase of family size by one person increases the log odds of having migrant by 0.131 indicating that large family size is one of the determinant factors for migration.

In conclusion, there is high prevalence of international migrant in the study area. The migration prevalence varies among the zones, the districts and the enumeration areas. Household head characteristics: age, educational level and occupation of head, and family size are determinant factors of international migration.

The district level random variation without and with six predictor variables, indicates there are 23.7% and 33% of the variation between districts and the rest 76.3% and 67% of variation are within the district correspondingly. The enumeration area level random variation without predictor variables in the model is 0.335 indicates that there is 33.5% of the variation between the enumeration areas and the 66.5% variation is within the enumeration area. Community based intervention is needed so as to monitor and regulate the international migration for the benefits of the society.

Acknowledgements

I would like to appreciate the professional guidance that I have received from my supervisor Dr. Ayele Taye and the administration supports from School of Mathematical and Statistical Sciences. I also acknowledge the HU-PhD Math Stat Project, Wachemo University and Hawassa University for their financial support during my study period.

References

Abrham, T., Sandra, A., and Aynadis, Y., (2014). Assessment on the Socio Economic Situation and Needs of Ethiopian Returnees from Kingdom of Saudi Arabia (KSA).

Berridge, D.M., and Crouchley, R. (2011). Multivariate Generalized Linear Mixed Models Using R. Lancaster University, by CRC Press, Taylor and Francis Group.

Browne W. J., Subramanian, S. V., Jones, K., and Goldstein, H. (2005). Variance Partitioning in Multilevel Logistic Models that Exhibit over Dispersion. *Journal of Royal Statistical Society*, 168, Part 3, pp. 599–613

Chi, G. and Voss, P. (2005). Migration Decision Making: A Hierarchical Regression Approach. *Journal of Regional Analysis and Policy*, 35: 11-22.

Cochran, W. G. (1977). Sampling Techniques. 3rd Edition. Harvard University, New York.

Faraway, J.J. (2006). Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models. Chapman & Hall.

Fay, R. E., and Herriot, R. A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. Journal of the American Statistical Association, Vol 74, 269-277.

Hilbe, J. (2009). Logistic Regression Models. Jet Propulsion Laboratory, California Institute of Technology and Arizona State University, U. S. A.

Hosmer, D. W., Jr., and S. Lemeshow (2000). Applied Logistic Regression. 2d edition. New York: John Wiley & Sons.

Jiang, J. (2007). Linear and Generalized Linear Mixed Models and Their Applications. Springer Series in Statistics.

Levy, P. S., and Lemeshow, S.(2008). Sampling of Populations: Methods and Applications, 4th edition.

Lohr, S.L. (2010). Sampling: Design and Analysis. 2nd Edition.

Longford, N.T. (2006). Sample Size Calculation for Small Area Estimation. Survey Methodology, 32(1):87-96.

McCulloch C.E, and Searle S.R. (2001). Generalized, Linear and Mixed Models, 2nd Edition. Wiley Publishing.

McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models, 2nd edition. Chapman and Hall.

Molefe, W.B. (2011). Sample Design for Small Area Estimation. Doctor of Philosophy Thesis, Center for Statistical and Survey Methodology, University of Wollongong, http://ro.uow.edu.au/theses/3495.

Naing L., Winn T., and Rusli B.N. (2006). Practical Issues in Calculating the Sample Size for Prevalence Studies. *Medical Statistics*, 1, 9-14.

Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. *Statistical Science*, Vol. 28, No. 1, 40–68. DOI: 10.1214/12-STS395.

Rango, M. and Laczko, F. (2014). Global Migration Trends: An Overview. International Organization for Migration, Saving Migrants Lives, Geneva.

Rao, J. N. K. (2003). Small Area Estimation. Wiley, Hoboken, NJ.

Setiawan, A. and Tarumi, T. (2004). Small Geographic Area Estimation in WinBUGS with Two Approaches Prediction. *Journal of the Faculty of Environmental Science and Technology*, Vo1. 9, No.1, pp. 9-17.

Teshome, D., Bailey. A, and Teller, Ch. (2013). Irregular Migration: Causes and Consequences of Young Adult Migration from Southern Ethiopia to Republic of South Africa. Paper Presented at the XXVII IUSSP International Population Conference 26-31 August, 2013 Busan, South Korea.

Thompson, M.E. (1997). Theory of Sample Surveys. Monographs on Statistics and Probability, 74, Department of Statistics and Actuarial Science University of Waterloo, Canada.

United Nations, Department of Economic and Social Affairs, Population Division, UNDESA (2016). International Migration Report 2015: Highlights. (ST/ESA/SER.A/375).

Zhao, Y., Staudenmayer, J., Coull, B.A., Wand, M.P. (2006). General Design Bayesian Generalized Linear Mixed Models. *Statistical Science*, Vol. 21, No. 1, 35–51.